(71) Applicants:
• Sony International (Europe) GmbH
10785 Berlin (DE)
• SONY CORPORATION
Tokyo 141 (JP)

(72) Inventors:
• Goronzy, Silke
c/o Sony Int. (Europe) GmbH
70736 Fellbach (DE)

• Kompe, Ralf
c/o Sony Int. (Europe) GmbH
70736 Fellbach (DE)
• Buchner, Peter
c/o Sony Int. (Europe) GmbH
70736 Fellbach (DE)
• Iwahashi, Naoto
Sony Computer Science Lab., Inc.
Shinagawa-ku, Tokyo 141-0022 (JP)

(74) Representative:
MÜLLER & HOFFMANN Patentanwälte
Innere Wiener Strasse 17
81667 München (DE)

(54) Semi-supervised speaker adaption

(57) To prevent adaptation to misrecognized words in unsupervised or on-line automatic speech recognition systems confidence measures are used or the user reaction is interpreted to decide whether a recognized phoneme, several phonemes, a word, several words or a whole utterance should be used for adaptation of the speaker independent model set to a speaker adapted model set or not and, in case an adaptation is executed, how strong the adaptation with this recognized utterance or part of this recognized utterance should be performed. Furtheron, a verification of the speaker adaptation performance is proposed to secure that the recognition rate never decreases (significantly), but only increases or stays at the same level.
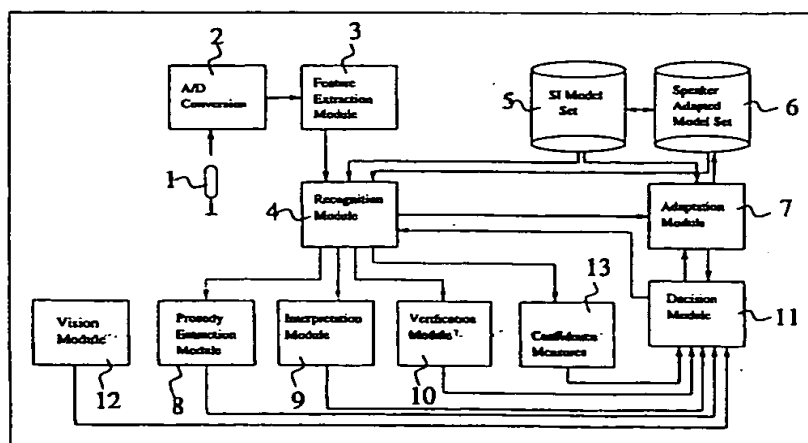
FIG. 1

EP 1 011 094 A1

Description

[0001]   This invention is related to automatic speech recognition (ASR), in particular to methods to perform an unsupervised or on-line adaption of an automatic speech recognition system and to a speech recognition system being able to carry out the inventive methods.

[0002]   State of the art speech recognizers consist of a set of statistical distribustions modeling the acoustic properties of certain speech segments. These acoustic properties are encoded in feature vectors. As an example, one Gaussian distribution can be taken for each phoneme. These distributions are attached to states. A (stochastic) state transition network (usually Hidden Markov Models) defines the probabilities for sequences of states and sequences of feature vectors. Passing a state consumes one feature vector covering a frame of e.g. 10 ms of the speech signal.

[0003]   The stochastic parameters of such a recognizer are trained using a large amount of speech data either from a single speaker yielding a speaker dependent (SD) system or from many speakers yielding a speaker independent (SI) system.

[0004]   Speaker adaptation (SA) is a widely used method to increase recognition rates of SI systems. State of the art speaker dependent systems yield much higher recognition rates than speaker independent systems. However, for many applications, it is not feasible to gather enough data from a single speaker to train the system. In case of a consumer device this might even not be wanted. To overcome this mismatch in recognition rates, speaker adaptation algorithms are widely used in order to achieve recognition rates that come close to speaker dependent systems, but only use a fraction of speaker dependent data compared to speaker dependent systems. These systems initially take speaker independent models that are then adapted so as to better match the speakers acoustics.

[0005]   Usually, the adaptation is performed supervised. That is, words spoken are known and the recognizer is forced to recognize them. Herewith a time alignment of the segment-specific distributions is achieved. The mismatch between the actual feature vectors and the parameters of the corresponding distribution builds the basis for the adaptation. The supervised adaptation requires an adaptation session to be done with every new speaker before he/she can actually use the recognizer.

[0006]   Figure 5 shows a block diagramm of such an exemplary speech recognition system according to the prior art. The spoken utterances received with a microphone 51 are converted into a digital signal in an A/D conversion stage 52 that is connected to a feature extraction module 53 in which a feature extraction is performed to obtain a feature vector e.g. every 10 ms. Such a feature vector is either used for training of a speech recognition system or after training it is used for adaptation of the initially speaker independent models and during use of the recognizer for the recognition of spoken utterances.

[0007]   For training, the feature extraction module 53 is connected to a training module 55 via the contacts a and c of a switch 54. The training module 55 of the exemplary speech recognition system working with Hidden Markov Models (HMMs) obtains a set of speaker independent (SI) HMMs. This is usually performed by the manufacturer of the automatic speech recognition device using a large data base comprising many different speakers.

[0008]   After the speech recognition system loades a set of SI models, the contacts a and b of the switch 54 are connected so that the feature vectors extracted by the feature extraction module 53 are fed into a recognition module 57 so that the system can be used by the customer and adaptated to him/her. The recognition module 57 then calculates a recognition result based on the extracted feature vectors and the speaker independent model set. During the adaptation to an individual speaker the recognition module 57 is connected to an adaptation module 58 that calculates a speaker adapted model set to be stored in a storage 59. In the future, the recognition module 57 calculates the recognition result based on the extracted feature vector and the speaker adapted module set. A further adapatation of the speaker adapted model set can be repeatedly performed to further improve the performance of the system for specific speakers. There are several existing methods for speaker adaptation, such as maximum a posteriori adaptation (MAP) or maximum likelihood linear regression (MLLR) adaptation.

[0009]   Usually, the speaker adaptation techniques modify the parameters of the Hidden Markov Models so that they better match the new speakers acoustics. As stated above, normally this is done in batch or off-line adaptation. This means that a speaker has to read a pre-defined text before he/she can use the system for recognition, which is then processed to do the adaptation. Once this is finished the system can be used for recognition. This mode is also called supervised adaptation, since the text was known to the system and a forced alignment of the corresponding speech signal to the models corresponding to the text is performed and used for adaptation.

[0010]   However, an unsupervised or on-line method is better suited for most kinds of consumer devices. In this case, adaptation takes place while the system is in use. The recognized utterance is used for adaptation and the modified models are used for recognizing the next utterance and so on. In this case the spoken text is not known to the system, but the word(s) that were recognized are taken instead.

[0011]   The EP 0 763 816 A2 proposes to use confidence measures as an optimization criterium for HMM training. These confidence measures are additional knowledge sources used for the classification of a recognition result as "probably correct" or "probably incor-

rect". Here, confidence measures are used for verification of n best recognized word strings and the result of this verification procedure. i.e. the derivative of the loss function, is used as optimization criterium for the training of the models. In this case, all utterances are used for training and the method is used to maximize the difference in the likelihood of confusable words. However, this document relates only to HMM training prior to system use.

[0012]   On the other hand, the EP 0 776 532 A2 discloses a method to correct misrecognition by uttering a predefined keyword "oops" whereafter the user might correct the misrecognized words by typing or the system tries to correct the error itself. In any case, the system only trains/adapts the speech models when a (series of) word(s) has been misrecognized.

[0013]   The present invention is concerned with the adaptation of speaker independent Hidden Markov Models in speech recognition systems using unsupervised or on-line adaptation. In these systems the HMMs have to be steadily refined after each new utterance or even after parts of utterances. Furtheron, the words that come into the system are not repeated several times and are not known to the system. Therefore, only an incremental speaker adaptation is possible. i.e. only very little adaptation data is available at a time. and additionally the problem arises that misrecognitions occur depending on the performance of the speaker independent system, because the output of the recognition module has to be assumed to be the correct word. These words are then used for adaptation and if the word was misrecognized, the adaptation algorithm will modify the models in a wrong way. The recognition performance might decrease drastically when this happens repeatedly.

[0014]   Therefore, it is the object underlying the present invention to propose a method and a device for unsupervised adaptation that overcome the problems described above in connection with the prior art.

[0015]   The inventive methods are defined in independent claims 1 and 17 and the inventive device is defined in independent claim 23. Preferred embodiments thereof are respectively defined in the following dependent claims.

[0016]   According to the invention, a kind of measurement indicates how reliable the recognition result was. The adaptation of the system is then based on the grade of the reliability of said recognition result. Therefore, this method according to the present invention is called semi-supervised speaker adaptation, since no supervising user or fixed set of vocabulary for adaptation is necessary.

[0017]   In case of a reliable recognition an utterance can be used for adaptation to a particular speaker, but in case of an unreliable recognition the utterance is discarded to avoid a wrong modification of the models. Alternatively, depending on the grade of the reliability a weight can be calculated that determines the strength of

the adaptation.

[0018]   The invention and its several methods of the decision whether to use an utterance for adaptation or not will be better understood from the following detailed description of exemplary embodiments thereof taken in conjunction with the appended drawings, wherein:

Fig. 1   shows a speech recognition system according to one embodiment of the present invention;

Fig. 2   shows a first adaptation method according to the present invention in which confidence measures are used;

Fig. 3   shows a second adaptation method according to the present invention in which a dialog history is observed;

Fig. 4   shows a method of switching back to the initial speaker independent models according to the present invention; and

Fig. 5   shows an exemplary speech recognition system according to the prior art.

[0019]   Fig. 2 shows a first adaptation method according to the present invention in which confidence measures are used to avoid adapting to a misrecognized word and to determine the grade of adaptation. This method is repeatedly executed in an endless loop beginning with step S21.

[0020]   In said first step S21 the recognition of a user utterance is performed like in a speech recognition system according to the prior art. In the following step S22 a confidence measurement is applied to the recognition result of step S21. In this step confidence measures are used to measure how reliable the recognition result is. In case the confidence measure is smaller than a certain threshold the recognized word is considered as unreliable and will not be used for adaptation so that the adaptation procedure is set forth again with step S21 in which the recognition of the next user utterance is performed. Is the confidence measure, on the other hand, above the threshold, the recognition result is considered to be reliable and used for adaptation in a step S23 before the adaptation procedure is again set forth with step S21 to recognize the next user utterance.

[0021]   To calculate a confidence measure according to the present invention, first one or several features are extracted from the recognition hypothesis and/or the speech signal. Then a decision is made based on these features whether the phoneme/word/phrase can be classified as correctly or incorrecetly recognized. This decision is no hard decision, but a certain probability for the correctness of a received utterance is calculated. The decision is e.g. based on a neural network or on decision trees which take the features as input and compute the confidence measure based upon some internal parameters.

[0022]   When a neural network is used to calculate the confidence measure, the output, i.e. the confidence

measure, is typically a value between 0 and 1; the closer this value is to 1, the more likely the phoneme/word/utterance or sequences thereof was recognized correct. Therefore, a threshold between 0 and 1 gets defined and confidence measures above said threshold classify a recognition result as correct.

[0023] The features based on which the confidence measure is computed are extracted from the recognition result or computed directly from the speech signal based on the recognition result. Such features can for example be the (relative) scores of the n-best recognition hypotheses, HMM state durations, durations of the recognized phonemes underlying the recognized words, or segment probabilities. The latter are computed by a stochastic model determining the probability for such a phoneme contained in a word hypothesis given an entire speech segment containing several frames.

[0024] The confidence measure can then directly be used to also determine the grade of adaptation. Of course, the simplest case of a confidence measure is to extract only one feature, e.g. the score provided by the HMMs during recognition, and to directly decide if the word was recognized correctly or not based on a threshold. In this case, the grade of adaptation is always constant.

[0025] As an alternative to the fixed threshold, the confidence measurement can be used to compute a weight which determines the strength of adaptation performed in step S23.

[0026] Furtheron, it is possible to change other parameters the decision is based on during adaptation, e.g. how the threshold used for deciding can be adapted depending on the derived features of the speech signal.

[0027] A problem occurs during speaker adaptation of the HMM models, because this influences the features of the confidence measure. This requires either a normalization of the features such that they are invariant to such changes of the HMM models or it requires an automatic on-line adaptation of the features or the parameters of the confidence measure or of the threshold to which the confidence measure is compared. This adaptation is based on a formal algorithm optimizing a criterion like the correctness of the confidence measure. The latter can be estimated based on the user reaction as determined in the vision, interpretation and prosody modules.

[0028] Furthermore, the confidence measure cannot only be applied to whole user utterances, but also word-wise or phoneme-wise so that not always the whole utterance is rejected for adaptation, but only the single misrecognized words or the words containing misrecognized phonemes. It is also possible to apply the confidence measures to a speech segment of another arbitrary length.

[0029] Such an adaptation guided by confidence measures needs no action from the user, such as announcing to the system that a word was misrecog-

nized. Therefore, this method achieves a considerably better recognition rate for unsupervised or on-line adaptation in automatic speech recognition systems than the systems according to the prior art, since not every user utterance or every word spoken by the user is used for adaptation irrespectively from the fact that such an utterance or word can be misrecognized and the grade of the adaptation depends on the probability of a correct recognized result.

[0030] Fig. 3 shows a second adaptation method according to the present invention in which the dialog history is observed to decide whether an utterance or single word or several words should be used for adaptation or not.

[0031] In a dialog system, the reaction of a user often shows if the recognized word was correct or not. A method to judge such a user reaction is shown in figure 3. Similar to the method depicted in figure 2, this method is repeatedly executed in an endless loop beginning with step S31.

[0032] In step S31 a recognition of a user utterance number i is performed like in the systems according to the prior art. Thereafter, the recognition result undergoes an interpretation in step S32 in which will be judged whether the user was satisfied with the systems reaction to his utterance spoken before the utterance number i. As an example such an utterance number i-1 could be "switch on the TV" and for some reasons the system recognized "switch on the radio" and thus the radio was switched on. When the user realizes this mistake, his/her next utterance (i.e. utterance number i) will be something like "no, not the radio, the TV" or "wrong. I said TV". In this case, the system will interpret in step S32 on basis of utterance number i that the previously recognized utterance was misrecognized and should not be used for adaptation. Step S33 in which the user utterance number i-1 is used for adaptation is in this case left out and step S34 in which the system performes an action or response is not carried out after step S33, but directly after step S32. After the action or response of the system in step S34, i is incremented in step S35 before the next utterance number i+1 of the user is recognized in step S31.

[0033] Apart from the wording or interpretation result of an utterance also information about the emotional state of a user, i.e. intonation and/or prosody, can be taken into account to judge whether the user is satisfied or not in step S32. So by interpreting the utterance using intonation and/or prosody, no special keywords are needed for the system to recognize that a misrecognition of the previously recognized utterance occured. For example, if a user says in an angry way to the system "turn on the TV" after his/her previously spoken utterance was misrecognized, the system can interpret that he/she did not change his/her mind, but that the previously recognized command was misrecognized so that it should not be used for adaptation.

[0034] Furtheron, also user reactions observed by

a visual computation system, such as a video camera connected to a computer that can interpret the reactions of a user, e.g. the mimic, can be used to verify a recognized utterance, e.g. based on a picture or a video sequence taken from the user and/or the user's face.

[0035] In this case it can be determined if the mimic shows anger or astonishment or if the lips of the user were closed although the recognizer recognized some words based on background voices or noise.

[0036] Depending on only one or a combination of those user reactions and on the intensity, a grade of adaptation can be determined. As in the case of confidence measures, it is also possible to set a threshold and therewith define a hard decision so that the grade of adaptation is constant.

[0037] Figure 4 shows a method according to the present invention, in which the system will switch back to the initial SI models, if the performance of the adapted models is too bad.

[0038] In this case, the system recognizes a situation in which adaptation was (repeatedly) done using misrecognized words, or a new speaker uses the system, since then the recognition performance may drop. Therefore, the system will switch back to the original speaker independent models. Similar to the methods depicted in figures 2 and 3, this method is repeatedly executed in an endless loop beginning with steps S41 and S43 that are executed in parallel.

[0039] Therefore, in said step S41 a recognition of a user utterance is performed using the adapted models, while in step S43 a recognition of the same user utterance is performed using the initial speaker independent models. To both recognition results a confidence measurement may be applied, respectively in steps S42 and S44. In a following step S45 both results, e.g. of the confidence measurements, are compared to decide whether to restart the adaptation with the initial speaker independent models in step S46 or to further use and adaptate the adapted models in a step S47, before the parallel recognition performed in steps 41 and 43 is performed with the next user utterance.

[0040] This method is not limited to the use of confidence measures to compare said both recognition results. It is also possible that the system uses other user reactions, e.g. of his/her behaviour of a certain time before and/or after the respective utterance or about intonation and/or prosody. It is also thinkable that th system asks the user to decide which models should be used, or which of the recognition results is the correct one and then use the respective model set for further recognition/adaptation.

[0041] Therefore, by keeping the original models and comparing their performance to the adapted ones, e.g. after a certain number of adaptation steps or in speech pauses, the initial models are also used by the system and adaptation is re-started in case the recognition result using the speaker independent models and/or the confidence measures indicate that the

adapted models do not perform as good as the initial ones. Therewith, it can be assured that the recognition rates never decrease (significantly), but only increase or stay at the same level. By performing this method the user's expectations are exactly satisfied, since a user would expect an automatic speech recognition system to get used to his way of speaking, just like humans do.

[0042] It is also possible that the speaker adapted models are not only compared to the speaker independent models to assure a recognition rate never decreasing (significantly), but also or instead to compare the newest speaker adapted models to older speaker adapted models to choose the ones having the best recognition performance and continue adaptation based on them.

[0043] Of course, all four methods according to the present invention described above or only a subset of them can be combined to prevent adaptation to misrecognized words or sentences in unsupervised or on-line adaptation mode. With these methods it is controlled whether adaptation is conducted with recognized words or a recognized utterance or not. Additionally a recognition rate never decreasing (significantly) is secured. As mentioned above, the proposed algorithms are independent from the adaptation methods themselves, i.e. they can be combined with any speaker adaptation algorithm.

[0044] An exemplary embodiment of a recognition system according to the present invention using either one or several of the inventive methods for unsupervised or on-line speaker adaptation is shown in figure 1.

[0045] In contrast to the speech recognition system according to the prior art shown in figure 5 the inventive system shown in figure 1 does not comprise a training module like the training module 55 of the prior art system or a similar circuit. This is no limitation of the system according to the present invention, since the training is performed independently of the adaptation with which the present invention is concerned. Of course, a switch provided behind the feature extraction module 3 to switch in-between adaptation/recognition mode and training mode, i.e. to lead the feature vectors either to the recogniton module 4, as it is shown in figure 1, or to a not shown training module which in turn can access the set of speaker independent modules that is stored in a storage 5 can also be provided.

[0046] Fig. 1 only shows the part of the automatic speech recognition system used for semi-supervised speaker adaptation according to the present invention. Therefore, the analog speech signal generated by a microphone 1 is converted into a digital signal in an A/D conversion stage 2 before a feature extraction is performed by a feature extraction module 3 to obtain a feature vector, e.g. every 10 ms. This feature vector is fed into a recognition module 4 that can access a storage 5 in which a speaker independent model set is stored, a storage 6 in which a speaker adapted model set is stored and an adaption module 7 that uses an adapta-

tion method. e.g. MAP or MLLR, to generate the speaker adapted model set by adaptation of the speaker independent model set. Therefore, the adaptation module 7 can access the speaker independent model set stored in the storage 5 via the storage 6 that is used for storing the speaker adapted model set. So far, all modules or storage devices are used in the same way like in the speech recognition system according to the prior art.

[0047]   According to the present invention, the recognition module distributes its results furtheron to a prosody extraction module 8 and an interpretation module 9 which perform both methods to decide whether a phoneme, several phonemes, a word, several words or a whole utterance should be used for adaptation or not as described above. Furtheron, the results of the recognition module is distributed to a confidence measure module 13 that calculates the confidence measures as described above. These modules lead their respective results to a decision unit 11 that decides whether adaptation is performed with said phoneme(s) single word, several words or whole utterances to provide its result to the adaptation module 7 which in turn uses this single phoneme(s), word, several words or whole utterance(s) to adapt the speaker adapted model set or not. The decision unit 11 also receives the output of a vision module 12 that represents the users visual behaviour corresponding to a certain utterance, i.e. his visual emotional state, e.g. if his mimic shows anger or astonishment, or if the user said something at all or if the recognized utterance was spoken by someone else.

[0048]   The decision whether the system should use speaker independent models or speaker adapted models is performed in a verification module 10 that receives both results of the recognition module 4, namely the result based on the speaker adapted model set and the result based on the speaker independent model set. The result of the verification module 10 influences the decision module 11, which passes also a control signal to the recognition module 4 determining which model set to use for the recognition and for the results passed to the prosody extraction module 8, the interpretation module 9 and the confidence measure module 13.

[0049]   Apart from changing the threshold to decide whether an utterance or part of an utterance should be used for adaptation, the input features of the decision module 11 can be adapted or also the parameters of the decision module 11 can be adapted.

[0050]   Of course, the decision unit 11 also determines the rate of the reliability of said single phoneme, several phonemes, single word, several words or whole utterance(s) to determine the strength of the adaptation that should be performed in the adaptation module 7. Also, the parameters used within the prosody extraction module 8, the interpretation module 9, the verification module 10 and the confidence measure module 13 can change dynamically as mentioned above. It is also pos-

sible that the decision module 11 does not switch immediatly to the speaker independent models, if they perform better, but waits some more utterances before this decision is made.

[0051]   Therefore, according to the preferred embodiment described above, the decision unit 11 receives the confidence measure (which parameters or features can be adaptive) of the spoken utterance or parts thereof, the information about prosody of the user when speaking the utterance or parts thereof, the interpretation of the user reaction determined on basis of the context of the spoken utterance, the interpretation of the visual user behaviour and the verification of the user to determine the grade of adaptation. Of course the invention is not limited to this and such a decision can also be based on a subset of this information.

Claims

1. Method to perform an unsupervised and/or on-line adaptation of an automatic speech recognition system, **characterized in that** a grade of adaptation of the system with the help of a received utterance or parts thereof is based on the grade of the reliability of the recognition result of said received utterance or the parts thereof.

2. Method according to claim 1, **characterized in that** said received utterance or a part of said received utterance is used for adaptation when the grade of the reliability of the recognition is above a threshold and is discarded when it is below said threshold.

3. Method according to claim 1 or 2, **characterized in that** said threshold is either fixed or dynamically changeable.

4. Method according to anyone of claims 1 to 3, **characterized in that** the grade of the reliability of the recognition result of said received utterance or a part of said received utterance is measured on the basis of confidence measures.

5. Method according to claim 4, **characterized in that** parameters and/or features said confidence measures are based on are adaptive.

6. Method according to claim 4 or 5, **characterized in that** said confidence measures are calculated on an utterance, word or phoneme based confidence score for each received utterance or part of said utterance.

7. Method according to claim 6, **characterized in that** said confidence score determines said grade of the reliability of the recognition result of said received utterance or a part of said received utterance.

8. Method according to anyone of claims 1 to 7, characterized in that the grade of the reliability of the recognition result of said received utterance or a part of said received utterance is measured on the basis of reactions of the speaker of said utterance.

9. Method according to claim 8, characterized in that said reactions are determined via a visual computation system based on a picture or a video sequence taken from the user and/or the user's face.

10. Method according to claim 8 or 9, characterized in that said confidence measures depend on the emotional state of the person speaking said utterance.

11. Method according to anyone of claims 8 to 10, characterized in that said reactions are determined by recognition and interpretation of utterances or parts of utterances received after said received utterance or parts of said received utterance.

12. Method according to claim 11, characterized in that said utterances or parts of utterances received after said received utterance or parts of said received utterance are checked for predefined keywords indicating that a previously received utterance was incorrectly or correctly recognized.

13. Method according to anyone of claims 10 to 14, characterized in that said reactions are determined by interpretation of secondary information of utterances or parts of utterances received after said received utterance or parts of said received utterance.

14. Method according to claim 13, characterized in that said secondary information of utterances or parts of utterances received after said received utterance or parts of said received utterance is intonation and/or prosody of said utterances or parts of utterances received after said received utterance or parts of said received utterance.

15. Method to perform an unsupervised or on-line adaptation of an automatic speech recognition system, in which adaptation of the system with the help of a received utterance or parts thereof is performed by repeatedly adapting a set of parameters, characterized in that at least one set of earlier parameters is stored to exchange the currently used parameters in case the recognition performance of the system drops.

16. Method according to claim 15, characterized in that the initial set of parameters is stored.

17. Method according to claim 15 or 16, characterized in that the recognition performance of the system is judged by comparing actual recognition results on the basis of stored earlier parameters and on the basis of the newest adapted parameters.

18. Method according to anyone of claims 15 to 17, characterized in that the recognition performance of the system is judged on the basis of the method defined in any of claims 1 to 17.

19. Method according to anyone of claims 1 to 28, characterized in that the adaptation of the system is performed using the adaptation of Hidden Markov Models.

20. Method according to claim 19, characterized in that it is used to adapt a speaker independent Hidden Markov Model towards the performance of a speaker dependent Hidden Markov Model.

21. Speech recognition system with unsupervised and/or on-line adaptation, comprising:

a microphone (1) to receive spoken words of a user and to output an analog signal;
an A/D conversion stage (2) connected to said microphone (1) to convert said analog signal into a digital signal;
a feature extraction module (3) connected to said A/D conversion stage (2) to extract feature vectors of said received words of the user from said digital signal;
a recognition module (4) connected to said feature extraction module (3) to recognize said received words of the user on basis of said feature vectors and a set of speaker independent and/or speaker adapted models;
an adaptation module (7) receiving the recognition result from said recognition module (4) to generate and/or adapt said speaker adapted model set;
characterized by
a decision unit (11) that is connected to said recognition module (4) and that supplies a signal to said adaptation module (7) indicating whether to use a certain received word for generation and/or adaptation of the speaker adapted model set or not.

22. Speech recognition system according to claim 21, characterized in that said signal supplied to said adaptation module (7) from said decision unit (11) indicates the strength of adaptation of the speaker adapted model set by said adaptation module (7) on basis of said certain received word.

23. Speech recognition system according to claim 21 or

7

22, **characterized in that** said signal supplied to said adaptation module (7) from said decision unit (11) is created on basis of a first control signal generated by a prosody extraction module (8) connected in-between said recognition module (4) and said decision unit (11).

5

24. Speech recognition system according to anyone of claims 21 to 23, **characterized in that** said signal supplied to said adaptation module (7) from said decision unit (11) is created on basis of a second control signal generated by a interpretation module (9) connected in-between said recognition module (4) and said decision unit (11).

10

15

25. Speech recognition system according to anyone of claims 21 to 24, **characterized in that** said signal supplied to said adaptation module (7) from said decision unit (11) is created on basis of a third control signal generated by a verification module (10) connected in-between said recognition module (4) and said decision unit (11).

20

26. Speech recognition system according to anyone of claims 21 to 24, **characterized in that** said signal supplied to said adaptation module (7) from said decision unit (11) is created on basis of a fourth control signal generated by a confidence measures module (12) connected in-between said recognition module (4) and said decision unit (11).

25

30

27. Speech recognition system according to anyone of claims 21 to 24, **characterized in that** said signal supplied to said adaptation module (7) from said decision unit (11) is created on basis of a fifth control signal generated by a vision module (12) connected to said decision unit (11).
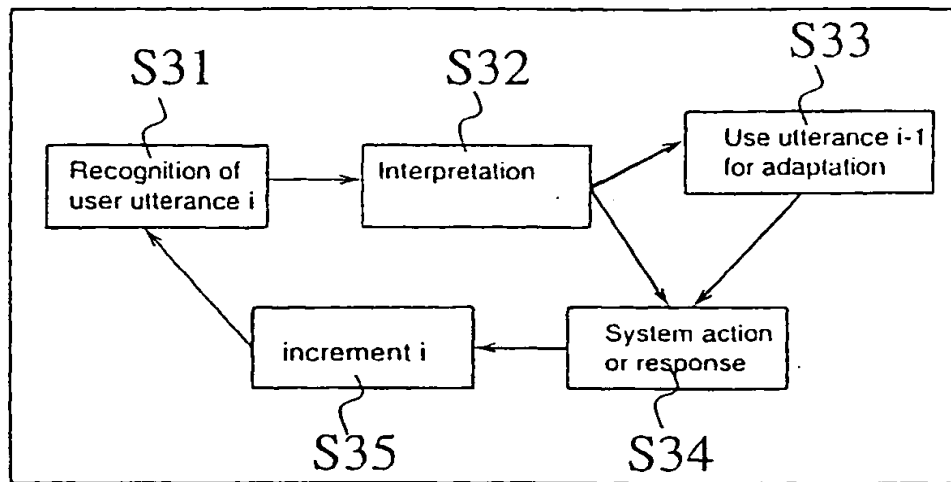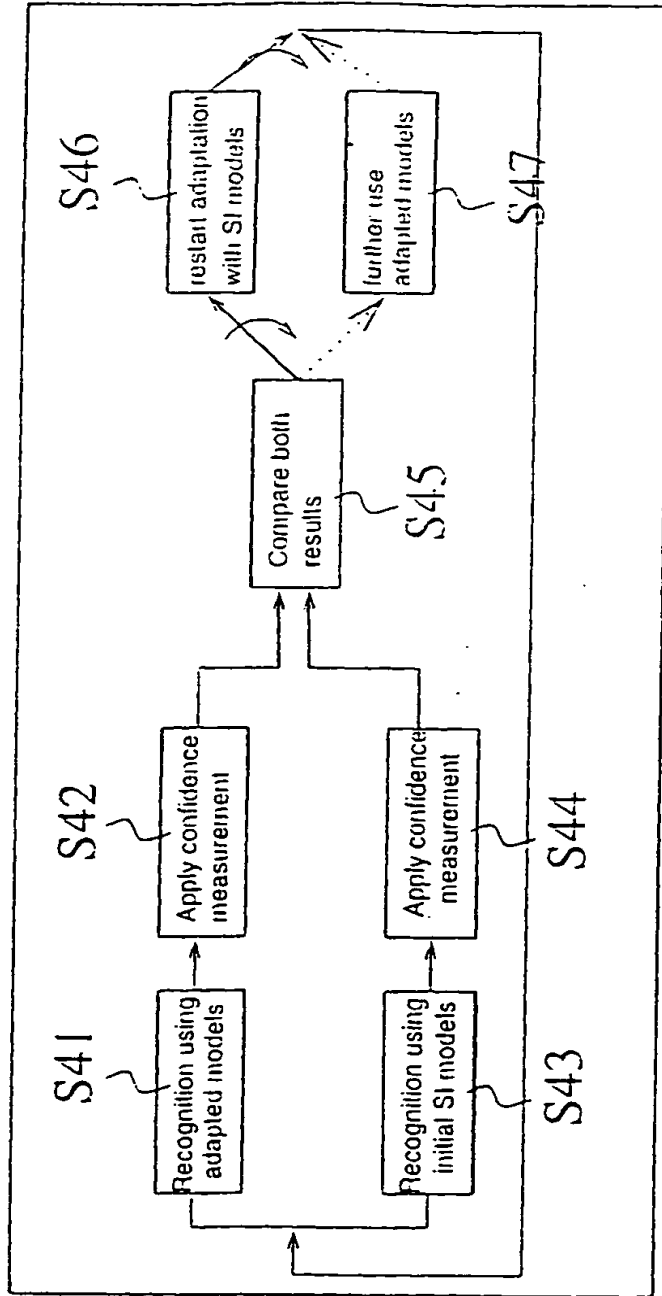
35

40

45

50

55
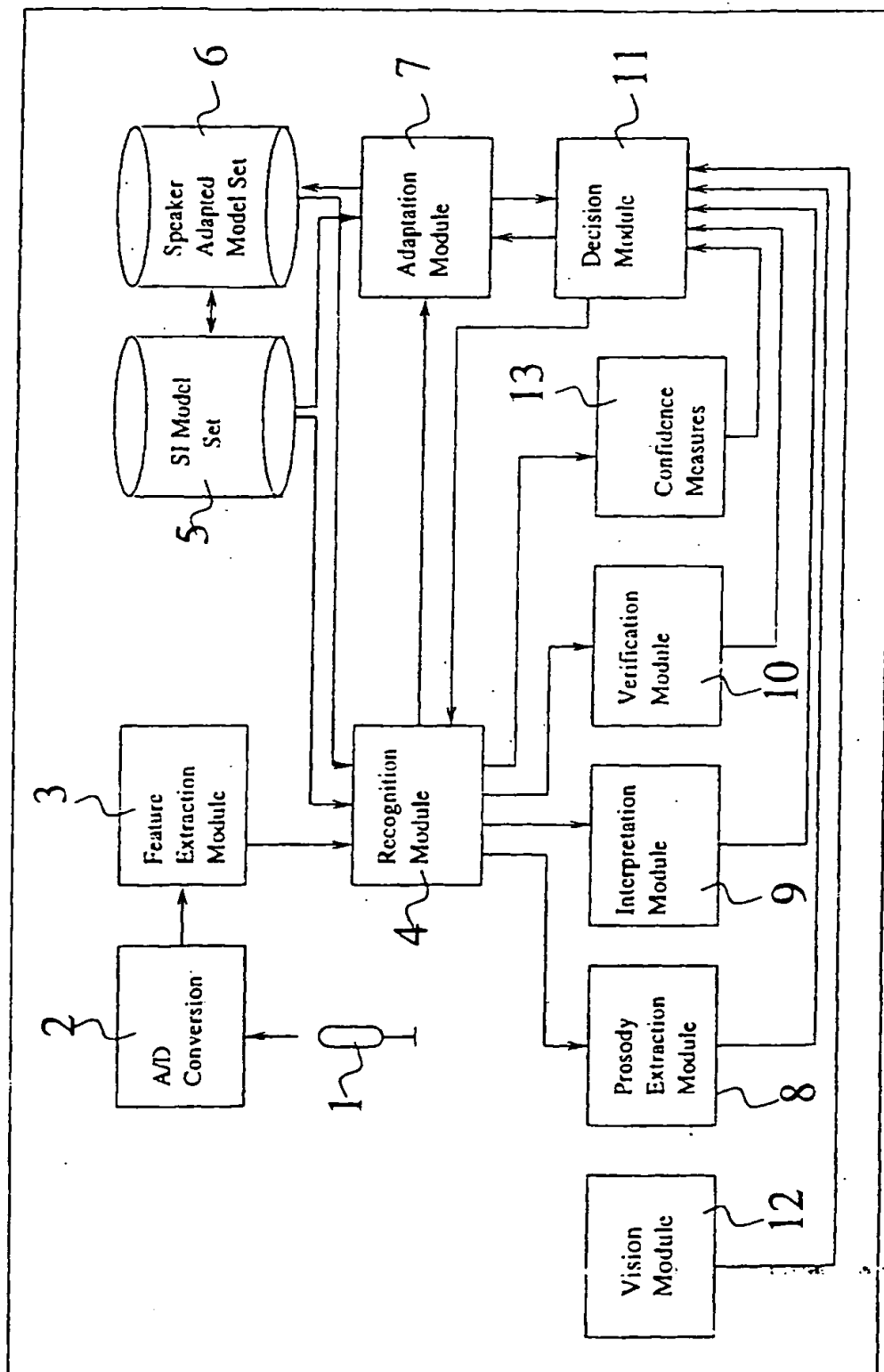
FIG. 1
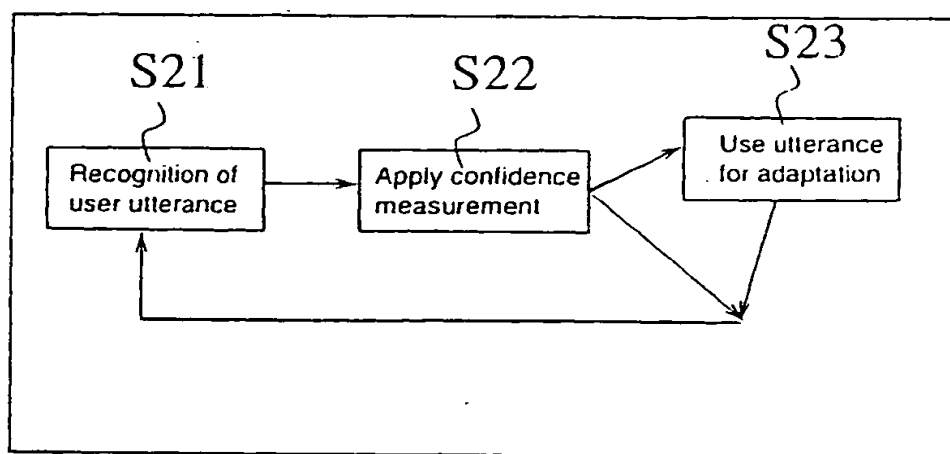
FIG. 2

FIG. 3

FIG. 4
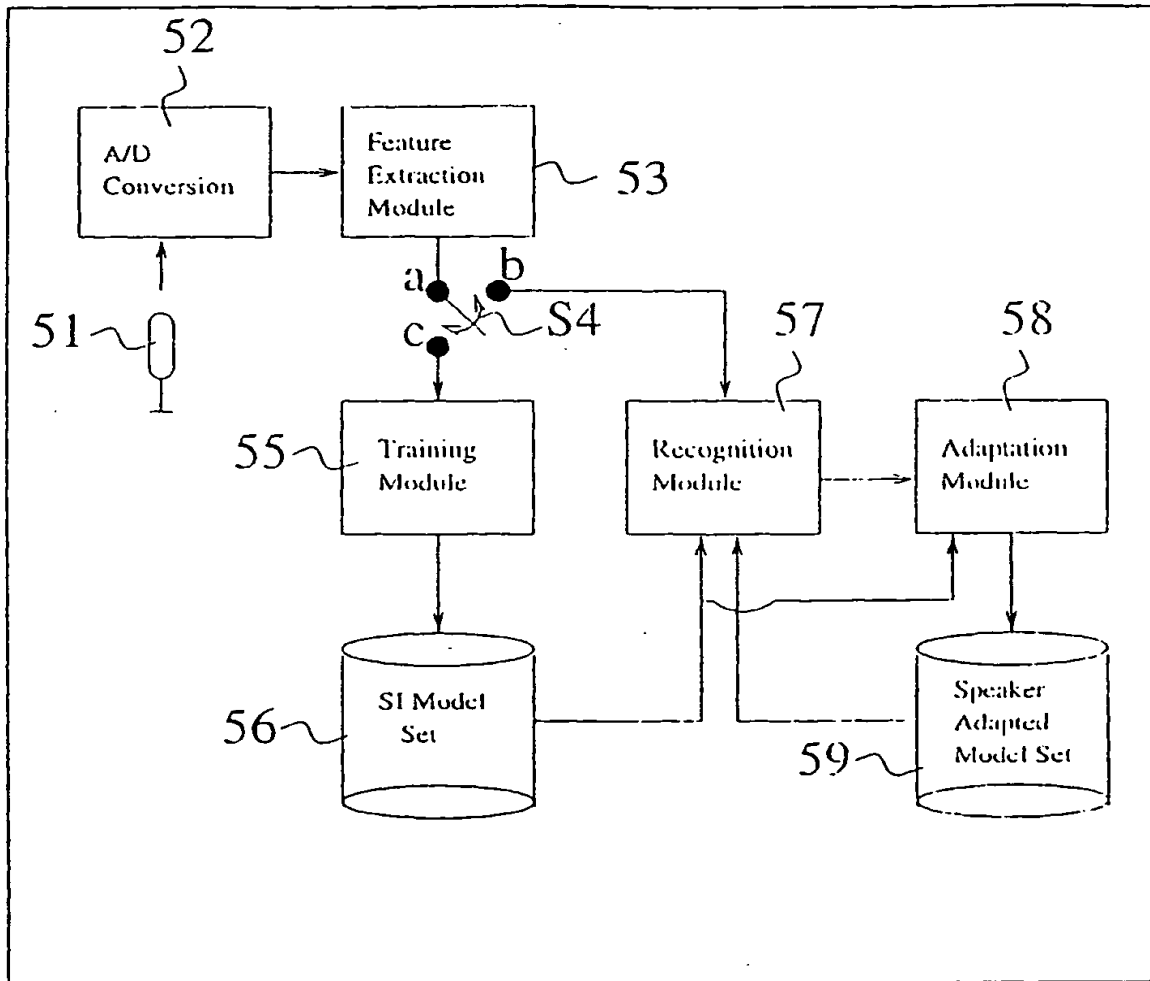
FIG. 5

European Patent Office

EUROPEAN SEARCH REPORT

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
X	HOMMA S ET AL: "Iterative unsupervised speaker adaptation for batch dictation" PROCEEDINGS ICSLP 96. FOURTH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (CAT. NO.96TH8206), PROCEEDING OF FOURTH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING. ICSLP '96, PHILADELPHIA, PA, USA, 3-6 OCT. 1996, pages 1141-1144 vol.2, XP002100688 ISBN 0-7803-3555-4, 1996, New York, NY, USA, IEEE, USA * paragraph 3 *	1-7, 19-22,25	G10L5/06
X	ZEPPENFELD T ET AL: "Recognition of conversational telephone speech using the JANUS speech engine" 1997 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (CAT. NO.97CB36052), 1997 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, MUNICH, GERMANY, 21-24 APRIL 1997, pages 1815-1818 vol.3, XP002100689 ISBN 0-8186-7919-0, 1997, Los Alamitos, CA, USA, IEEE Comput. Soc. Press, USA * paragraph 5 *	1-7,21, 22,25	
A	EP 0 660 304 A (AT & T CORP) 28 June 1995 * abstract *	8,10	TECHNICAL FIELDS SEARCHED (Int.Cl.6) G10L
	-/--		

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	24 June 1999	Krembel, L.

EUROPEAN SEARCH REPORT

		European Patent Office			Application Number
					EP 98 12 4024

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
X	US 5 842 168 A (INAZUMI MITSUHIRO ET AL) 24 November 1998 * column 3, line 9 - line 20 * * column 3, line 37 - line 50 * * column 4, line 42 - line 58 * * column 9, line 6 - line 9 * * column 10, line 60 - column 11, line 10 * * figure 1 * ---	15-17	
A	US 5 144 672 A (KURIKI SHOJI) 1 September 1992 * abstract * * column 1, line 14 - line 29 * * figure 1 * -----	15-18	
			TECHNICAL FIELDS SEARCHED (Int.Cl.6)

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	24 June 1999	Krembel, L.

EPO FORM 1503 03 82 (P04C01)

15

European Patent
Office

CLAIMS INCURRING FEES

The present European patent application comprised at the time of filing more than ten claims.

☐ Only part of the claims have been paid within the prescribed time limit. The present European search report has been drawn up for the first ten claims and for those claims for which claims fees have been paid, namely claim(s):

☐ No claims fees have been paid within the prescribed time limit. The present European search report has been drawn up for the first ten claims.

LACK OF UNITY OF INVENTION

The Search Division considers that the present European patent application does not comply with the requirements of unity of invention and relates to several inventions or groups of inventions. namely:

see sheet B

☒ All further search fees have been paid within the fixed time limit. The present European search report has been drawn up for all claims.

☐ As all searchable claims could be searched without effort justifying an additional fee, the Search Division did not invite payment of any additional fee.

☐ Only part of the further search fees have been paid within the fixed time limit. The present European search report has been drawn up for those parts of the European patent application which relate to the inventions in respect of which search fees have been paid, namely claims:

☐ None of the further search fees have been paid within the fixed time limit. The present European search report has been drawn up for those parts of the European patent application which relate to the invention first mentioned in the claims, namely claims:

European Patent Office

LACK OF UNITY OF INVENTION
SHEET B

Application Number

EP 98 12 4024

The Search Division considers that the present European patent application does not comply with the requirements of unity of invention and relates to several inventions or groups of inventions. namely:

1. Claims: 1-14,19-27

   Conditional speaker adaptation

2. Claims: 15-18

   Selection of the best model for speech recognition

ANNEX TO THE EUROPEAN SEARCH REPORT
ON EUROPEAN PATENT APPLICATION NO.

EP 98 12 4024

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report.
The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

24-06-1999

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0660304 A	28-06-1995	US 5559926 A	24-09-1996
		AU 689151 B	26-03-1998
		AU 8045394 A	29-06-1995
		CA 2137301 A,C	23-06-1995
		JP 2798620 B	17-09-1998
		JP 7199988 A	04-08-1995
US 5842168 A	24-11-1998	JP 9062287 A	07-03-1997
US 5144672 A	01-09-1992	DE 4031638 A	18-04-1991
		JP 3206500 A	09-09-1991

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

BNSDOCID: <EP___1011094A1_I_>